09/705450

cg c

PATENT
Attorney Docket No. 39209

## IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent No. 7,012,889

Confirmation No. 7885

Issued: March 14, 2006

Name of Patentee: TURNER ET AL.

Patent Title: METHOD AND APPARATUS FOR CONTROLLING INPUT RATES WITHIN A PACKET SWITCHING SYSTEM

**CERTIFICATE OF MAILING**

I hereby certify that this paper is being deposited with the United States Postal Service on the date shown with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, Washington, D.C. 20231, on **August 21, 2006.**

Kirk D. Williams, Esq. 8-21-06

## REQUEST FOR CERTIFICATE OF CORRECTION OF PATENT FOR PATENT OFFICE MISTAKE (37 C.F.R. § 1.322)

Attn: Certificate of Correction Branch
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

It is requested that a Certificate of Correction be issued to correct Office mistakes found the above-identified patent. Attached hereto is a Certificate of Correction which indicates the requested correction. For your convenience, also attached are copies of selected pages (a) from the issued patent with errors highlighted, and (b) from the original application as filed November 2, 2000, (c) Amendment A filed February 16, 2005, and (d) the Notice of Allowance mailed June 16, 2005 with the correct text/instructions.

In re US Patent No. 7,012,889

It is believed that there is no charge for this request because applicant or applicants were not responsible for such error, as will be apparent upon a comparison of the issued patent with the application as filed or amended. However, the Assistant Commissioner is hereby authorized to charge any fee that may be required to Deposit Account No. 501430.

Respectfully submitted,
**The Law Office of Kirk D. Williams**

Date: August 21, 2006

By 8-21-2006
Kirk D. Williams, Reg. No. 42,229
One of the Attorneys for Applicants
CUSTOMER NUMBER 26327
The Law Office of Kirk D. Williams
1234 S. OGDEN ST., Denver, CO 80210
303-282-0151 (telephone), 303-778-0748 (facsimile)

# UNITED STATES PATENT AND TRADEMARK OFFICE
# CERTIFICATE OF CORRECTION

PATENT NO. : 7,012,889

DATED : March 14, 2006

INVENTOR(S) : Turner et al.

It is certified that error(s) appear in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

Col. 3, line 56, replace "time to" with -- time $t_0$ --

Col. 3, line 67, replace "setting to" with -- setting $t_0$ --

Col. 12, line 66, replace "claim 5" with -- claim 4 --

Col. 13, line 1, replace "claim 5" with -- claim 4 --

Col 13, line 6, replace "claim 6" with -- claim 7 --

MAILING ADDRESS OF SENDER:

Kirk D. Williams, Reg. No. 42,229
Customer No. 26327
The Law Office of Kirk D. Williams
1234 S. Ogden Street, Denver, CO 80210

PATENT NO. 7,012,889

No. of additional copies

⇨ NONE (0)

environments, and embodiments in keeping with the scope and spirit of the invention. Some embodiments described may include, inter alia, systems, integrated circuit chips, methods, and computer-readable media containing instructions.

A system is described that includes rate monitors that measure the rate at which traffic arrives for each of the outputs of the system and includes a set of rate-controlled virtual output queues at each input line card. In one embodiment, there is one queue for each output of the system. Some embodiments further include a rate assignment mechanism that determines what rates should be assigned to each of the rate-controlled virtual output queues. These rate-controlled virtual output queues also include a mechanism for adjusting the rates at which packets are sent to the outputs of the system. These rate-controlled virtual output queues may include the mechanism for automatically determining and adjusting their sending rates, or receive this information from another source (e.g., another component, external source, etc.). In one embodiment, these sending rates are adjusted based on received flow control information.

The system receives flow control information corresponding to the status of each of the outputs of the system. In one embodiment, the system includes an interconnection network that maintains separate internal buffers for each of the different output links and sends XON and XOFF flow control signals to the input ports as necessary to regulate the flow of packets to different outputs. The ability to control input rates within a system is not limited to any particular flow control scheme. Numerous mechanisms are known in the art for accumulating and distributing flow control information in systems, including those for use in packet switching and other communications systems, and therefore, are not discussed with particularity herein.

In one embodiment, a rate monitor M(i,j) for traffic from input i to output j includes a state machine S(i,j) with three states: unconstrained, off and backlogged. If output j is not congested (e.g., the total traffic going to output j does not exceed the bandwidth of the interface to the output line card) then S(i,j) is unconstrained. S(i,j) goes to the off state whenever the input line card at input i receives a flow control signal turning off traffic to output j. S(i,j) goes from the off state to the backlogged state whenever it receives a flow control signal turning on traffic to output j. S(i,j) goes from the backlogged state to the unconstrained state when the queue at input i for output j becomes empty.

In one embodiment, when S(i,j) is unconstrained (e.g., the output is not congested), packets are sent to output j at their arrival rate. When S(i,j) is off (e.g., the output is in a off state), packets are not sent to output j. When S(i,j) is backlogged (e.g., the output is in a backlogged state), packets are sent to output j at a reduced pacing rate approximately proportional to their arrival rate.

$t_0$

In one embodiment, the rate at which traffic arrives for congested outputs is monitored. One method of doing this is to keep a record of the last time to when the queue at input i for output j was empty and to count the number of packets, c, received since time $t_0$. A measured average arrival rate, R(i,j), at time t is then equal to $c/(t-t_0)$. The pacing rate is then set according to the formula, pacing rate=f*R(i,j), where f is a parameter of the system and is called the acceleration factor. An alternative to measuring the average arrival time from the last time the queue was empty is to measure the average arrival time during successive measurement intervals while the queue remains non-empty. This can be done, for example, by clearing c periodically and at the same time setting to equal to the current time. This approach allows the pacing rate to more quickly adapt to changes in the rate at which traffic arrives. In other embodiments, the pacing rate is determined with additional parameters. For example, in systems which support packets of varying lengths, the pacing rate may be based on the size of the received packets (e.g., total bytes, etc.), rather than, or in addition to a count of packets.

$t_0$

Different embodiments employ various acceleration factors f, which may substantially vary between different systems. Acceleration factor f may be set at system configuration time or may be varied during the operation of the system based on some parameters, such as traffic congestion. In one embodiment, acceleration factor f is related to the speed-up factor of the packet switching fabric over the packet arrival rate. For example, in one embodiment system having a speed-up factor of 1.3, an acceleration factor f of approximately 1.2 is used.

In one embodiment, each input i has a queue for each output and a queue scheduler that determines when packets are sent from each queue. At any point in time, a queue at input i for a backlogged output j is assigned a rate P(i,j) and the queue scheduler sends packets to output j at the assigned rate, whenever S(i,j)=backlogged (when S(i,j)=off, no packets are sent from input i to output j).

Let T(i,j)=1/P(i,j) be the target time interval between successive packets sent from input i to output j. T(i,j) is expressed in units equal to the time it takes an input line card to send a packet to the interconnection network.

In one embodiment, the queue scheduler is a data structure that comprises a set of "timing wheels." A timing wheel can be implemented as a one-dimensional array of linked lists. Each list contains a set of queue identifiers. The position of a list in the array is used to determine when the queue so identified should next send a packet to the output link. In the simplest case, a single timing wheel is used. In such an embodiment, indicators of outputs are stored in the timing wheel data structure until their scheduled time. At this time, the indicators are removed from the timing wheel data structure and placed in a transmit list. Items are removed from the transmit list and a packet corresponding to the output is sent, with an indicator for the output re-inserted into the timing wheel data structure in an appropriate time bin if packets remain to be sent to the output.

The time bin into which a queue identifier is inserted, is selected to produce the desired rate of transmission from that queue. For each queue, there is a parameter T(i,j) referred to as the inter-packet time for that queue. This parameter gives the average number of packet times between successive cell transmissions from the queue. To enable accurate rate specifications, the inter-packet time may be expressed in time units that are smaller than the time it takes to transmit a single packet. When a queue identifier is re-inserted into a time bin, a target transmission time is computed for the next packet to be sent from that queue. This target transmission time is equal to T(i,j) plus the target transmission time of the previous packet sent from the queue. The queue identifier is re-inserted into that time bin whose contents will be transferred to the transmit list at the time that is closest to the target transmission time.

In one embodiment, each timing wheel also has a cursor which points to one of the lists in the array. The cursors are advanced from one position in the array to the next position in the array as time advances. The cursor for the first timing wheel is advanced at every time step (a time step being the time it takes an input line card to send a packet to the interconnection network). The cursor for the second timing wheel is advanced less frequently, the cursor for the third wheel **601** is maintained with the current time indicated by cursor **602**. A transmit list **604** is also maintained to indicate outputs which are allowed to be sent a packet, and in which order. In the illustrated embodiment, timing wheel **601** and transmit list **604** both use linked list data structures and include output queue identifier elements **603A** and **603B** (which may be in the form of output queue identifier data structure **500** illustrated in FIG. 5A).

At the current time indicated by cursor **602**, output queue identifier elements **605** are moved from timing wheel **601** to the tail of transmit list **604**. In parallel, the output queue identifier element **606** at the head of transmit list **604** is removed and a corresponding packet, stored in a packet queue (not shown) is sent to the corresponding output. If the output is in the "BACKLOGGED" state, the output queue identifier element **606** is rescheduled and placed in timing wheel **601** at an appropriate place corresponding to a next time to send the next packet to the corresponding output. In one embodiment, this next time is proportional to the measured and maintained average packet arrival rate for the output as previously discussed herein.

One embodiment for maintaining the state of an output in response to received flow control information is illustrated in the flow diagram of FIG. 7A. Processing begins at process block **700** and proceeds to process block **705**, where flow control information is received for an output. Next, as determined in process block **710**, if the output's current state is "UNCONSTRAINED," then if an XOFF flow control signal is received as determined in process block **712**, then the output's state is set to "OFF" in process block **714**, and the packet count for the output is reset in process block **716**.

Otherwise, as determined in process block **720**, if the output's current state is "OFF," then if an XON flow control signal is received as determined in process block **722**, then if the output's output queue is empty as determined in process block **730**, then the output's state is set to "UNCONSTRAINED" in process block **732**. Otherwise, the output's state is set to "BACKLOGGED" in process block **734**, and an output queue identifier corresponding to the output is placed in the transmit list in process block **736**.

Otherwise, the output is in the "BACKLOGGED" state, and as determined in process block **742**, if an XOFF flow control signal is received, then the output's state is set to "OFF" in process block **744**.

Processing then returns to process block **705** to receive more flow control information.

The operation of one embodiment in response to a received packet is illustrated in FIG. 7B. Processing begins at process block **755**, and proceeds to process block **760** where a packet destined for a particular output is received. Next, in process block **765**, the received packet is placed in an output queue corresponding to the output destination of the received packet. Next, as determined in process block **770**, if the current state of the output is "UNCONSTRAINED," then an output queue identifier is placed in the transmit list in process block **772**.

Otherwise, the packet count is increased for the output in process block **775**. Then, as determined in process block **780**, if the output's current state is "BACKLOGGED," then if the output is not already scheduled in the transmit list as determined in process block **790**, then an output queue identifier is placed at the end of the transmit list in process block **795**.

Processing then returns to process block **760** to receive more packets.

The operation of an embodiment for processing the transmit list is illustrated in FIG. 8. Processing begins at process block **800**, and proceeds to process block **805**, where an output queue identifier is removed from the head of the transmit list. Next, in process block **810**, a packet is retrieved from the head of the indicated output queue and sent to the output. Next, as determined in process block **815**, if the output's state is "BACKLOGGED", then, if the output queue corresponding to the output just sent a packet is empty as determined in process block **820**, then the output's state is set to "UNCONSTRAINED" in process block **825**. Otherwise, the output queue identifier is rescheduled in process block **830**.

Processing then returns to process block **805** to send more packets.

For simplicity of understanding, some embodiments have been described herein using one type of data structures and/or elements. Typically, these data structures and elements have been described in the form of a linked list. As is apparent to one skilled in the art, numerous other embodiments are possible which use one or more of a wide variety of data structures and elements in keeping with the scope and spirit of the invention.

In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes maybe made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

What is claimed is:

1. An apparatus comprising:
   - a plurality of rate monitors to measure the rate at which traffic arrives for each of a plurality of outputs of a packet switching system;
   - one or more state data structures indicating a state of each of the plurality of outputs of the packet switching system; and
   - a rate-controlled virtual output queue for each of the plurality of outputs of the packet switching system, each of the rate controlled virtual output queues adjusting a rate at which packets are sent to a particular destination based at least in part on the measured traffic arrival rate for the particular destination and the state for the particular destination;
   - wherein the one or more state data structures maintains an indication of one of at least three different states for each of the plurality of outputs of the packet switching system; and
   - wherein packets are not sent to a particular output when the particular output is in a first state, packets are sent to the particular output at approximately the measured traffic arrival rate when the particular output is in a second state, and packets are sent to the particular output at a reduced rate approximately proportional to the measured traffic arrival rate when the particular output is in a third state.

2. An input line card comprising the apparatus of claim 1.

3. The apparatus of claim 1, wherein each of the rate-controlled virtual output queues includes a transmit list.

4. The apparatus of claim 1, wherein each rate-controlled virtual output queue includes a timing mechanism.

5. The apparatus of claim 1, wherein each of the plurality of rate monitors include one or more data structures maintaining an indication of a packet count and a reference time period.

4

6. The apparatus of claim 5, wherein the timing mechanism includes one or more timing wheels.

claim 4

Depends from Issued claim 4

7. The apparatus of claim ~~5~~, wherein the rate-controlled virtual output queue comprises at least one scheduling data structure, said at least one scheduling data structure including scheduling information with a timing granularity greater than that of the timing mechanism. 7

Depends from Issued claim 7

8. The apparatus of claim ~~6~~, wherein the one or more state data structures maintains an indication of one of at least three different states for each of the plurality of outputs of the packet switching system.

9. The apparatus of claim 7, wherein said scheduling information includes a target time for sending a next packet.

10. A method performed by a packet switching system, the method comprising:

receiving packets at a first component of the packet switching system, at least a subset of the received packets being destined for a second component of the packet switching system;

maintaining a state data structure indicating a state of the second component;

maintaining a rate data structure reflective of an arrival rate at which packets destined for the second component are received at the first component;

sending received packets to the second component at a first rate approximately proportional to the arrival rate when the state data structure indicates the second component is in a first state; and

sending received packets to the second component at a second rate less than the first rate and greater than zero, and approximately proportional to the arrival rate when the state data structure indicates the second component is in a second state.

11. The method of claim 10, wherein the first rate is approximately the arrival rate of the received packets.

12. The method of claim 10, wherein the rate data structure includes a count of a subset of the received packets.

13. The method of claim 10, wherein a set of possible states for the state of the second component includes an unconstrained state, an off state, and a backlogged state.

14. The method of claim 13, further comprising sending no received packets to the second component from the first component when the state data structure indicates the second component is in an off state.

15. A method performed by a packet switching system, the method comprising:

receiving a plurality of packets, each of the received plurality of packets being destined for one or more of a plurality of outputs of the packet switching system;

measuring a traffic arrival rate for each one of the plurality of outputs of the packet switching system, the traffic arrival rate reflective of the rate at which traffic arrives for a corresponding one of the plurality of outputs of the packet switching system;

maintaining an indication of a state of said each one of the plurality of outputs of the packet switching system;

sending received packets to a particular one of the plurality of outputs at a first rate approximately proportional to the measured traffic arrival rate for the particular one of the plurality of outputs when the maintained state indication reflects the particular one of the plurality of outputs is in a first state; and

sending received packets to the particular one of the plurality of outputs at a second rate less than the first rate and greater than zero, and approximately proportional to the measured traffic arrival rate for the particular one of the plurality of outputs when the maintained state indication reflects the particular one of the plurality of outputs is in a second state.

16. The method of claim 15, wherein no packets are sent to a particular one of the plurality of outputs when the maintained state indication reflects the particular one of the plurality of outputs is in a third state.

17. The method of claim 15, wherein said indications of said states of the plurality of outputs are updated based on received flow control information.

18. The method of claim 15, wherein said method is performed by an input line card of the packet switching system.

19. The method of claim 15, wherein measuring the traffic arrival rate includes maintaining a packet count and a time reference.

20. The method of claim 15, further comprising:

maintaining a packet queue for each output of the packet switching system; and

placing each packet of the plurality of received packets in one of the plurality of packet queues based on a destination of said each packet.

21. The method of claim 20, further comprising placing an indicator of a corresponding one of the plurality of packet queues in a transmit list upon arrival of a particular received packet having a destination of a selected one of the plurality of outputs being in the first state.

22. The method of claim 15, wherein sending received packets to the particular one of the plurality of outputs at the second rate includes:

sending one of the plurality of packets to the particular one of the plurality of outputs of the packet switching system; and

rescheduling the particular one of the plurality of outputs of the packet switching system in a timing data structure for a second scheduled time based upon the measured traffic arrival rate for the selected output.

23. The method of claim 22, wherein sending received packets to the particular one of the plurality of outputs at the second rate includes retrieving a transmit indication corresponding to the particular one of the plurality of outputs of the packet switching system from the timing data structure at a first scheduled time.

24. The method of claim 22, wherein the second scheduled time reflects an actual time to send one of the plurality of packets to the selected output of the packet switching system rather than a time relative to a last sent packet to the selected output of the packet switching system.

25. The method of claim 22, wherein the timing data structure includes one or more timing wheels.

26. The method of claim 22, comprising maintaining a target time for the sending one of the plurality of packets, wherein the second scheduled time is approximately the target time.

27. The method of claim 26, wherein the target time has a finer timing resolution than that of the timing data structure.

28. The method of claim 15, wherein sending received packets to the particular one of the plurality of outputs at the second rate includes:

retrieving a transmit indication corresponding to a selected output of the plurality of outputs of the packet switching system from a timing data structure at a first scheduled time and placing the retrieved transmit indication in a transmit list;

removing the retrieved transmit indication from the transmit list and sending one of the plurality of packets to the corresponding selected output of the plurality of outputs of the packet switching system based on the retrieved transmit indication; and

29 2000

From Application Filed November 2, 2000

those for use in packet switching and other communications systems, and therefore, are not discussed with particularity herein.

In one embodiment, a rate monitor M(i,j) for traffic from input i to output j includes a state machine S(i,j) with three states: unconstrained, off and backlogged. If output j is not congested (e.g., the total traffic going to output j does not exceed the bandwidth of the interface to the output line card) then S(i,j) is unconstrained. S(i,j) goes to the off state whenever the input line card at input i receives a flow control signal turning off traffic to output j. S(i,j) goes from the off state to the backlogged state whenever it receives a flow control signal turning on traffic to output j. S(i,j) goes from the backlogged state to the unconstrained state when the queue at input i for output j becomes empty.

In one embodiment, when S(i,j) is unconstrained (e.g., the output is not congested), packets are sent to output j at their arrival rate. When S(i,j) is off (e.g., the output is in a off state), packets are not sent to output j. When S(i,j) is backlogged (e.g., the output is in a backlogged state), packets are sent to output j at a reduced pacing rate approximately proportional to their arrival rate.

In one embodiment, the rate at which traffic arrives for congested outputs is monitored. One method of doing this is to keep a record of the last time $t_0$ when the queue at input i for output j was empty and to count the number of packets, c, received since time $t_0$. A measured average arrival rate, R(i,j), at time t is then equal to $c/(t-t_0)$. The pacing rate is then set according to the formula, pacing rate=f*R(i,j), where f is a parameter of the system and is called the acceleration factor. An alternative to measuring the average arrival time from the last time the queue was empty is to measure the average arrival time during successive measurement intervals while the queue remains non-empty. This can be done, for example, by clearing c periodically and at the same time setting $t_0$ equal to the current time. This approach allows the pacing rate to more quickly adapt to changes in the rate at which traffic arrives. In other embodiments, the pacing rate is determined with additional parameters. For example, in systems which support packets

From Amendment A filed February 16, 2005

In re TURNER ET AL., Application No. 09/705,450
Amendment A

**Amendments to the Claims:**

The listing of clams will replace all prior versions, and listings, of claims in the application:

**Listing of Claims:**

Claims 1 (canceled)

Claim 2 (currently amended): An input line card comprising the apparatus of ~~claim 1~~ claim 4.

Issued as Claim 8 Claim 7

Claim 3 (currently amended): The apparatus of ~~claim 1~~ claim ~~8~~, wherein the one or more state data structures maintains an indication of one of at least three different states for each of the plurality of outputs of the packet switching system.

From Amendment A filed February 16, 2005

Claim 4 (currently amended): ~~The apparatus of claim 3,~~ An apparatus comprising:

a plurality of rate monitors to measure the rate at which traffic arrives for each of a plurality of outputs of a packet switching system;

one or more state data structures indicating a state of each of the plurality of outputs of the packet switching system; and

a rate-controlled virtual output queue for each of the plurality of outputs of the packet switching system, each of the rate controlled virtual output queues adjusting a rate at which packets are sent to a particular destination based at least in part on a measured traffic arrival rate and a state for the particular destination;

wherein the one or more state data structures maintains an indication of one of at least three different states for each of the plurality of outputs of the packet switching system; and

wherein packets are not sent to a particular output when the particular output is in a first state, packets are sent to the particular output at approximately the measured traffic arrival rate when the particular output is in a second state, and packets are sent to the particular output at a reduced rate approximately proportional to the measured traffic arrival rate when the particular output is in a third state.

Claim 5 (currently amended): The apparatus of ~~claim 1~~ claim 4, wherein each of the rate-controlled virtual output queues includes a transmit list.

Issued as Claim 4

~~Claim 6~~ (currently amended): The apparatus of ~~claim 1~~ claim 4, wherein each rate-controlled virtual output queue includes a timing mechanism.

FROM Notice of Allowance mailed 6-16-2005

**EXAMINER'S AMENDMENT**

1. An examiner's amendment to the record appears below. Should the changes and/or additions be unacceptable to applicant, an amendment may be filed as provided by 37 CFR 1.312. To ensure consideration of such an amendment, it MUST be submitted no later than the payment of the issue fee.

Authorization for this examiner's amendment was given in a telephone interview with Kirk D. Williams on 6/9/2005.

The application has been amended as follows:

In the claims:

In lines 8-9 of claim 4, "a measured traffic arrival rate and a state for the particular destination" has been corrected to --the measured traffic arrival rate for the particular destination and the state for the particular destination--.

Replace claim 7 with --~~Claim 7~~ Issued as claim 6: The apparatus of ~~claim 6~~ claim 4, wherein the timing mechanism includes one or more timing wheels.--

Replace claim 8 with --~~Claim 8~~ Issued as claim 7: The apparatus of ~~claim 6~~ claim 4, wherein the rate-controlled virtual output queue comprises at least one scheduling data structure, said at least one scheduling data structure including scheduling information with a timing granularity greater than that of the timing mechanism.--

2. The following is an examiner's statement of reasons for allowance:

To further clarify the reasons for allowance, the prior art of record fails to disclose, in combination with the other limitations of the independent claims, the combination of the